有關係？沒關係？——談迴歸與相關

◎農試所農藝系　呂秀英

一、前　言

當試驗研究的目的在探知兩特性值x與y間的相互關係時，我們通常會使用到迴歸(regression)與相關(correlation)之統計分析方法。人們對迴歸與相關常混淆不清，在科學文獻上相關的問題常被當做迴歸的問題來處理，而迴歸的問題也常被當做相關的問題來處理。造成這種混淆不清的原因有數種。首先這兩種分析間的數學關係很密切，在數學上我們可以很容易地由一種關係轉到另一種關係。其次，是因為早期的教科書對這兩種方法間的區分沒有解釋清楚，其所造成的困擾到現在還無法完全被克服。而且至少還有一本現在的教科書把其中的一種分析當做是另一種分析的異名，增加更多的困擾。最後，研究者所選用的方法雖然可能正確，可是所要分析的數據可能不適於使用該法來分析。茲謹就這些問題進行討論。

二、到底應該用迴歸或相關分析？

由於迴歸係數和相關係數之間有一些簡單的數學關係（兩者的計算都要用到乘積和），因此，計算迴歸係數時，常會使人聯想到要計算相對的相關關係；但是這樣做是不正確的，除非我們在研究的開始就打算探知互相關聯的程度且所得到的數據也適用於這種計算。有時對一個問題我們不易分清到底是應該用迴歸分析或是用相關分析。其實很簡單，只要從分析目的及數據性質來判斷，就可以很容易予以清楚區分。

首先討論這二種分析的目的之不同。迴歸分析是想描述一個依變數(dependent variable)y對一個自變數(independent variable)x的倚賴情形。在這裡，x可以利用各種試驗設計方法來自由變動，例如生長箱溫度、土壤含氮量、施肥用量、作物苗齡、殺蟲劑濃度等等。我們利用迴歸方程式來支持x的改變產生y的改變之假說，並可用來預估某一x的y值，以及將x當做是統計控制來解釋y的一些變異，例如乾物重受溫度之影響程度、生長速率受施肥用量之影響程度、昆蟲族群致死率受殺蟲劑濃度之影響程度等。

相反地，相關分析是涉及兩個變數的相互倚賴之程度，即在一起變異的程度。我們並不把一個變數表示成另一個變數的函數，所以沒有自變數和依變數之分。也就是說，x為不可自由變動，例如預測颱風或探求水

作者：呂研究員秀英
聯絡電話：(04)3302301轉125

稻穗數與精米重間、雜草種子數目與成長日數間、乳牛體重與泌乳量間的關係等問題時，雖然有可能一個變數是另一個變數的因，但我們既不知道也不假設有這兩種關係存在。通常，這兩個變數是同時受到某一共同原因的影響（但未必一定是如此）。當我們想測定一個全體的成對變數間之互相關聯程度時，適宜採用相關分析。

即使我們選用的方法符合我們的目的，但是數據之性質可能不適合使用這種方法來處理，例如想將穀粒產量當做是穗數的一個函數，我們可能取同一期作同一試區內的一個植株逢機樣本，而得到每一株的穀粒產量與穗數而做前者對後者的迴歸。可是，這兩個變數之測量都有誤差。自變數x的各別測量值並非由試驗者隨意選擇或控制，因此傳統迴歸的假定不成立，用此數據做迴歸是不合理的。這種錯誤，不難從發表過的研究文獻找到例子。

難道我們眞的不可以用一個方程式來描述y對x的倚賴情形？如果眞的想，就必須使用不同模式架構的迴歸。傳統迴歸，是假定x爲可以自由選擇或控制的固定值，這種模式屬於模式I迴歸，方程式中的迴歸係數是利用最小平方法(least squares method)來估計，一般我們都是使用模式I迴歸；若自變數x不是固定值，就必須採用模式II迴歸，由於兩變數同時有逢機變異，所以方程式中的迴歸係數是利用最大概似法(maximum likelihood method)來估計。模式II迴歸的計算非常複雜，使用較不普遍。

三、從迴歸或相關分析下正確的結論

雖然迴歸分析通常被用來檢測因果關係，但是y對x的一個顯著迴歸並不證明x的改變是y之變異的原因。縱然模式II迴歸可以用來解釋兩個變數一起變異的程度，但是，還是最好不要將變異和因果混爲一談。不同的變數可能一起變異，但這種共變異可以只是偶發現象或可能同時受到一種相同的原因所支配。因此，在解釋迴歸分析的結果時，無論是模式I或模式II迴歸，我們若發現操縱一個變數會影響第二個變數時，也只能說自變數x的變異會引發依變數y的變異（但是不能說就是該變數的因！有時候其他教科書會把y翻譯成因變數，但我們卻比較喜歡用依變數的中譯名，也是基於這個理由），但是即使這樣說，也得非常小心。例如我們發現某種作物的生長速率是其周圍溫度的一個函數，我們可以下結論說溫度是造成該作物生長速率改變的原因之一。但必須注意，可能還有其他因素影響速率。一個容易犯的錯誤，是把因果關係反過來說。當然不會有人認爲作物生長速率會影響周圍環境的溫度，但是我們卻可能會對植株兩種生長性狀間的因果關係下錯誤的結論。當我們發現y對x的迴歸不顯著時，除了非常複雜的情形以外，我們可以很有把握地說x的離差不會影響y。

相關分析是用來測量任何一對變數間互相關聯的程度，並測驗這個程度是否大於單靠機遇所可期望得到的程度。如果相關被確定，則可進一步研究導致變數間的相關之眞正原因。我們要強調不要把顯著的相關和因果關係混爲一談。而且也要當心所謂的「無意義相關」，一個衆所皆知的無意義相關的例子是在美國超過一萬居民的城市之每人平均酒消耗量與該城市的牧師之人數間呈正相關。農業上無意義相關的例子也不少，例如

單株產量與公頃產量間呈正相關、地上部與地下部乾重分配比例間呈負相關、調查性狀在不同取樣時間之間呈正相關、昆蟲族群數在夏季增大與雜草高度間呈正相關等等。每一個相關的情形都必須加以仔細分析，有的相關可能是由於一個變數是另一個變數的部份或全部之因，例如公頃產量是單株產量的導出性狀，它們之間本來就存在有比例關係；例如地上部與地下部的乾重分配比例總和為1（100％），當然進行其間的相關分析，一定呈負相關；因此，進行這些相關分析一點意義也沒有。各時間點所取樣的性狀資料，本來其間就具有相關性，並非獨立，即每一個時間的表現都會受到前段時間的影響，故與其進行相關分析，倒不如探討這些性狀隨時間的變化趨勢，可以獲得更多的資訊。但是有的相關則是因為兩個變數同時受到某一個原因的影響或由於更複雜的情況所造成，例如居民酒消耗量與牧師人數間的關係其實是由城市的大小所造成，而昆蟲族群在夏季的增大，有些雜草的高度也增高，其間的相關可能只是因為時間的改變所造成，而未必有生態上的關係。

相關係數的顯著性會受到樣本數很大的影響。除非所取的樣本非常大且相關係數非常接近於1或-1時，我們可以毫無疑問認定兩者間有明顯的直線關係傾向，否則最好還是將兩個變數成對的資料，製成分佈圖或相關圖，以瞭解兩個變數之間的關係，如此可以避免分析判斷上的錯誤。特別是相關係數的值不大但顯著的情況，更應該在下結論前先繪成分佈圖。樣本數太少，常是造成相關係數值不大但卻顯著的原因之一，當然這樣的分析結果的準確性便不值得信賴。當樣本是取自於多個異質族群時，各群內x與y間無關係存在，但從群平均值看之，x與y卻具有比例關係，這種群間異質性是無法從顯著的相關係數察覺出來的，但是從分佈圖就可以一目了然。

採用適當的樣本大小，經過統計檢定而發現不顯著的相關時，人們很容易會作出兩個變數間無顯著關係的結論，但這不是很正確的說法。當相關係數顯著時，我們通常解釋成兩變數間呈正(負)相關。但嚴格來講，該兩變數的關係，指的是直線（一次）關係。相關分析，其實是簡單相關的簡稱，目的在測定y與x之間是否有最簡單的直線關係。因此，不顯著的相關，最多也只能說是兩個變數之間沒有顯著的「直線」關係，卻不能證明其間就沒有其他更複雜的非線性關係。

四、決定係數的平方根等於相關係數？

對迴歸分析而言，決定係數(coefficient of determination, R^2)是一個很有用的統計量。在迴歸分析中是利用變方分析來把總平方和細分為已被解釋與未被解釋的平方和。做這種變方分析之後，可以得到已被解釋平方和除以總平方和的比值，當做是總變異因迴歸而被解釋之比例的一種估計。比例愈接近於1，表示迴歸方程式愈有效。

在直線迴歸分析時，R^2恰好為簡單相關係數r的平方值。但這純粹是一種數學關係，而且只發生在直線迴歸。複迴歸的決定係數之開方根，就不等於簡單相關係數。在複迴歸分析中，將R^2開方，可以得到另一種統計量－複相關係數(multiple correlation coefficient)，通常我們是以R的符號來表示，以便與簡單相關係數r作一區分。R和R^2都是

用來表示迴歸方程式的有效性，二者選其一即可。但是因為人們常錯誤地將複相關係數(R)與相關係數(r)混為一談，因此我們通常是採用R^2。

決定係數（複相關係數）與相關係數的目的不同，適用的數據性質也不一樣。如同迴歸，決定係數（複相關係數）適用於y逢機而x固定的資料場合；但相關係數則適於y與x皆逢機的變數性質。因此，不能把迴歸分析中的複相關係數，當做是相關係數。常看到很多科學文獻在迴歸方程式之後，一起列上相關係數，這是不正確的。因為適用於模式I迴歸的數據性質，絕不適用於相關分析。除非我們是採用模式II迴歸，但幾乎很少人會用到這個方法。在迴歸方程式之後，應該列出決定係數，而不是相關係數。

五、結　語

相關係數的應用已有一段很長的歷史，近年來由於生物學的研究逐漸變為試驗研究，所以相關的應用稍為減少。在試驗中通常是將一個因素以人為控制來檢查另一變數的反應，因此較適宜採用迴歸的方法。但是無論在生物學或其他科學，仍然有些問題不能用試驗的方法來解決，因為有些變數無法受研究者的控制。尤其在生態學、系統學、演化學、社會學等科學領域，氣候、演化歷史、人類心理等之各種因素無法被人所改變。但是，就像生物化學或實驗胚胎學的問題一樣，我們也必須要瞭解這些現象的科學機制。此時，相關分析可被先用來描述變數間的互相關聯程度。

眞實的相關通常被相信是因果的關連，但無論是採用迴歸或相關，最好不要將變異和因果混為一談。不同的變數可能一起變異，但是這種共變異可以只是偶發現象或可能同時受到一種相同的原因所支配。因此，最正確的說法是，迴歸是用來測量依變數y受到自變數x的影響程度，而相關是用來探知兩變數（無依變數及自變數之分）間的互相關聯程度。迴歸的x為可以自由變動的固定值，而相關的x是不可自由變動的逢機變數。